CNN Based Framework for Sentiment Analysis of Tweets

Vikas Tripathi, Bhasker Pant, Vijay Kumar

Abstract*: The project deals with the problem of visual updates on twitter; that differentiates tweets according to the context in which they are exposed: good or bad. Twitter is an online small-scale platform for showcasing different thoughts perceptions related to any area, news, information etc. It is an informal communication framework that allows clients to compose brief information of 280 characters long. It's a quickly developing assistance with more than 400 million enlisted clients of which 326 million individuals are dynamic and half of them sign on twitter each day - creating almost 500 million tweets every day. Considering this huge measure of spending we want to pick up the declaration of open assessment by examining the feelings communicated in the tweets. Investigating general assessment is fundamental for any kind of business. firms trying to identify the appropriate response of their items in the market, foreseeing political decisions and anticipating financial occasions, for example, stock costs. The purpose of this project is to develop an effective working class for accurate and automated segmentation of the tweet stream.

***Keywords*: CNN, Tweets, Sentiment analysis, Machine Learning.**

I. INTRODUCTION

With the ongoing development of versatile data frameworks and the expanding accessibility of advanced cells, online networking has become a significant piece of regular daily existence in numerous networks. This improvement has included the advancement of enormous volumes of information [1]: information that when broke down can be utilized to remove significant data about different subjects. Assumption investigation (SA), otherwise called idea mining is the way toward ordering feelings passed on by a book, for instance as negative or positive. Information made available on social media contributes to the fragmentation of the research work within SA in recent times and to the shifting focus of the field to this type of data [2]. The information obtained from using SA on social media data has many features that can be used, for example, to help marketers analyze the success of an ad campaign, to identify how different people have received product releases, to predict user performance, or to predict election results. The most popular social networking site is Twitter, 1 small blogging site that allows users to write up to 280-character posts, referred to as tweets [3]. As of November 2018, Twitter has 326 million monthly active users on their homepage, with approximately 88% of their free-to-read tweets. In addition, more than 84% of users have their site specified on their profiles, which enables them to automatically scroll to locations. Twitter-generated data is made available through the Twitter API and represents a real-time stream of the data you have. Tweets can be filtered both by location and time published [4]. This has opened the way to a new field of SA: Twitter (TSA) psychology [5]. The purpose of this project is to distinguish the given tweets, which are introduced as large information utilizing Machine Learning calculations [6]. AI is a sort of man-made brainpower (AI) that enables PCs to learn without being unequivocally customized. AI centers essentially around the improvement of PC frameworks that can change when new information is introduced.

Revised Manuscript Received on March 20, 2019.

***** Correspondence Author

1. Dr. Vikas Tripathi*, CSE, Graphic Era deemed to be university, Dehradun, India, vikastripathi.be@gmail.com

2. Dr. Bhasker Pant, CSE, Graphic Era deemed to be university, Dehradun, India, pantbhaskar2@gmail.com

3. Dr. **Vijay Kumar,** Department of Physics, Graphic Era Hill University, Dehradun.

II. LITERATURE REVIEW

The conceptual analysis of the domain of small-scale blogging is a relatively new topic of research so there is still much room for further research in this area [7]. A decent amount of previous work related has been done in user reviews of user reviews, documents, blogs / web articles and standard analysis [8]This differs from twitter mainly due to the 280 character limit per tweet that forces the user to express a compressed view in a very short text. Good results have been achieved in the use of segmentation using guided learning techniques such as Naive Bayes, Vector Support Machines and the Convolutional Neural Network but the manual writing of the supervised method is very expensive.

III. METHODOLOGY

Our proposed project is about classification of tweets using python. This is achieved by using many types of features from the tweets and then applying basic classification techniques like SVM (Support Vector Machine), Decision Tree, Naïve Bayes and CNN (Convolution Neural Network). CNN architecture is shown in fig. 1. These classification techniques are able to classify the points in separate folder by working with the respective features we provide to them. After the classification, we compare the accuracy of different methods of for sentiment analysis. First of a large dataset of tweets in English language is collected from different open source databases then preprocessing of data is done to convert data into desired format so that the model training and testing on the basis of dataset becomes precise and accurate.

After the collection of suitable data, we need to process the data. In this phase all the unwanted characters are removed from the tweets, such as special symbols, abbreviations, hashtags, usernames etc. and the tweets are divided into two files, one with positive tweets and the other with negative tweets. After this, with the help of these two files we create a

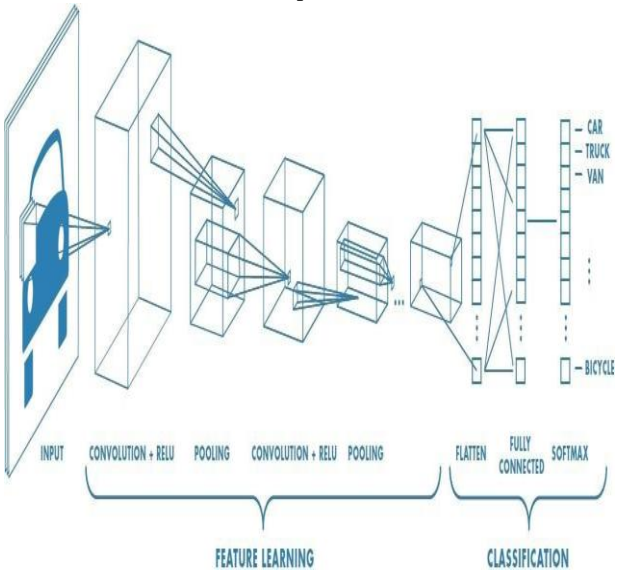

Fig. 1 CNN architecture

vocabulary which contains all the unique words which is contained in data repository. This word data set is separated into two primary parts known as train and test. In training data is fed for machine leering to acquire the knowledge for prediction. Once training gets completed words fed to check which kind of prediction can be obtained from specific algorithm. In context of this paper analysis have been performed for identification of tweets having positive intent or negative intent. For example, when filtering emails "spam" or "not spam", when looking at transaction data, "fraudulent", or "authorized".

IV. RESULTS AND DISCUSSIONS

We have divided dataset in 90:10 (Training: Testing) and performed classification on dataset using different classifiers whose results are given below: The data set was processed before building a CNN model. As we know that there is a limit of 280 characters and as a result, users are forced to use different abbreviations, special symbols etc., for ex- #NaMo, @RaGa and so as to clean all such expressions from the tweets, a pre-processing of data is need to be completed before anyone could actually build a CNN model.

Table I Accuracy achieved by various algorithms

S. No.	Classifier	Accuracy (%age)
1.	Naïve Bayes	60.4
2.	Decision Tree	65.9
3.	SVM	71.42
4.	CNN	80.2

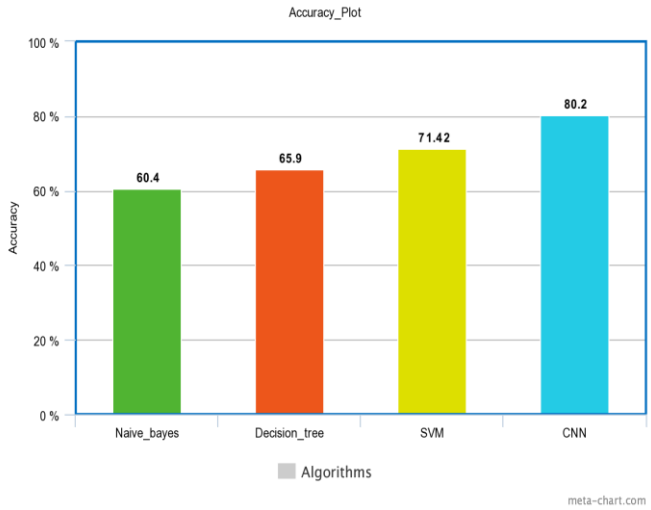

Fig. 2 Graphical representation of accuracies achieved by various algorithms

Later we used different fractions of dataset so as to train the CNN model and the results are as follows-

From the table 2 and fig. 2, we can clearly observe that Convolution Neural Network (CNN) gives the best results on the dataset. The accuracy achieved by CNN is approximately 80.2% which very high as compare to other classifiers applied to the dataset.

Table II Accuracy description using different fraction of dataset

S. No.	Algorithm	Dataset Fraction	Accuracy
1.	CNN	50%	78.14%
2.		66%	79.55%
3.		75%	80.27%

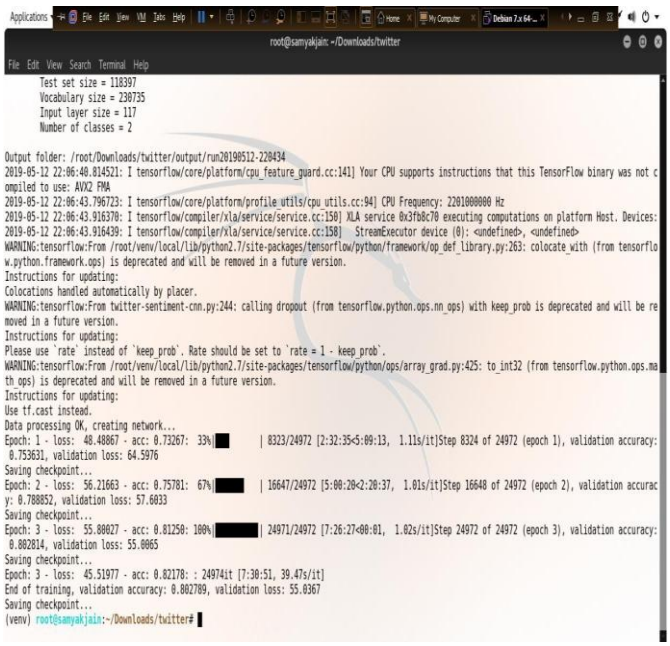

Fig. 3 CNN Model output generation

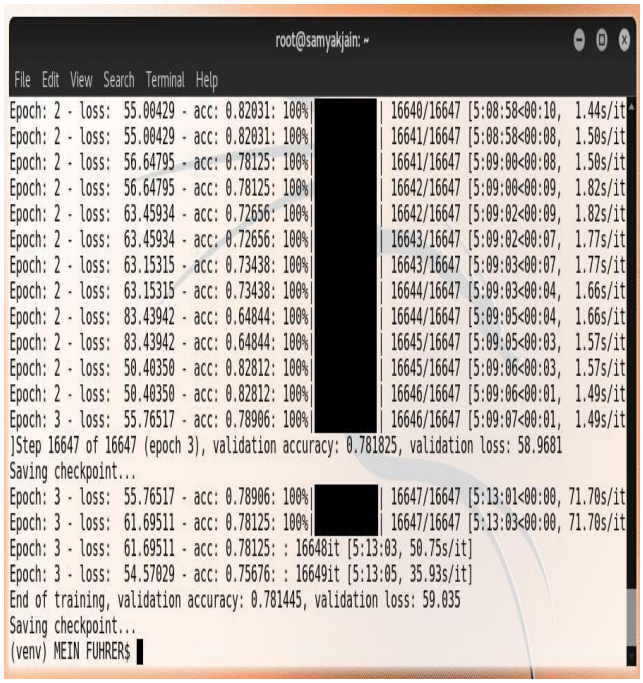

Fig. 4 CNN processing when data is 50 %

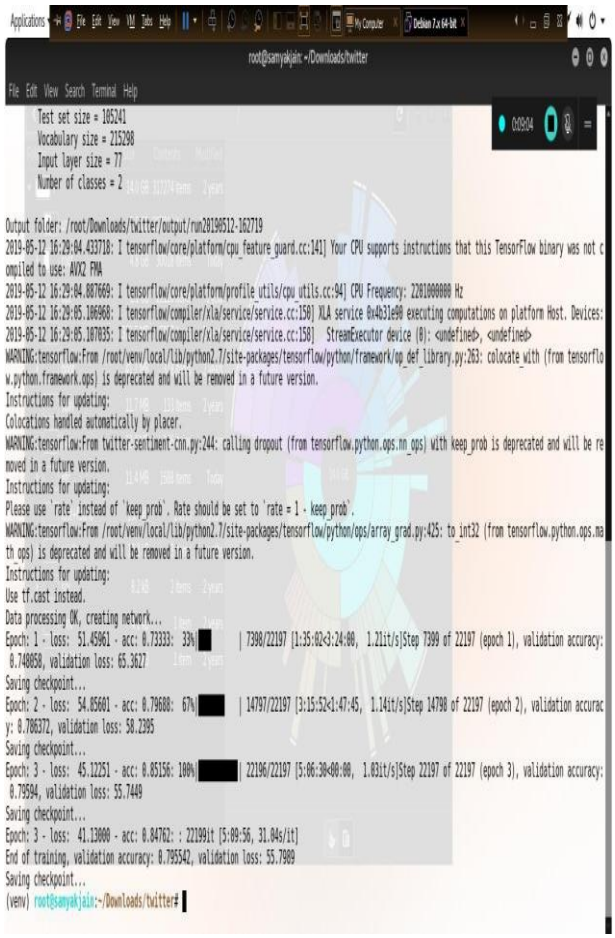

Fig. 5 Data process using CNN

V. CONCLUSION

Analysis of feelings/ sentiments of any individual, especially in the domain of micro-blogging, is still have huge potential for researchers. Right now, by using a single layer CNN model, in our experiments, we achieved a maximum accuracy of 80.2%. However, using a multiple layer CNN the accuracy of the model may further increase. Also, there could be many other Machine Learning algorithms which could be used so as to improve the accuracy of sentiment analysis, also the dataset can be enhanced not only in terms of quantity but also in terms of quality, i.e. more information about the tweets could be added to the dataset.

REFERENCES

1. M. Grothaus. (2018, Oct. 25). Twitter's Q3 earnings by the numbers [online]. Available: https://www.fastcompany.com/90256723/twitters-q3-earnings-by-the numbers
2. S. Aslam. (2019, Jan. 6). Twitter by The Numbers: Stats, Demographics and Fun Facts [online]. Available: https://www.omnicoreagency.com/twitter-statistics/
3. M. Grothaus. (2018, Oct. 25). Twitter's Q3 earnings by the numbers [online]. Available: https://www.fastcompany.com/90256723/twitters-q3-earnings-by-thenumbers
4. D. Britz. (2015, Dec. 11). Implementing a CNN for Text Classification in TensorFlow [online]. Available: http://www.wildml.com/2015/12/implementing-a-cnnfor-text-classification-in-tensorflow/
5. D. Britz. (2015, Nov. 7). Understanding Convolutional Neural Network for NLP [online]. Available: http://www.wildml.com/2015/11/understanding-convolutionalneural-networks-for-nlp/
6. D. Gupta. (2017, June. 29). Architecture of Convolutional Neural Network Demystified [online]. Available: https://www.analyticsvidhya.com/blog/2017/0 6/architecture- of-convolutional- neuralnetworks-simplified-demystified/
7. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
8. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification.

AUTHORS PROFILE

Dr. Vikas Tripathi has done BE in information technology from Technocrats institute of technology, Bhopal, M. Tech in Software engineering from Indian institute of information technology Gwalior and PhD from Uttarakhand technical university, Dehradun. He is actively involved in research related to Software engineering, Computer Vision, Machine learning and Video Analytics. He has published many papers in reputed international conferences and journals. Currently he is working as an associate professor in Graphic era deemed to be university Dehradun, India..

Dr. Bhaskar Pant Currently working as Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal. Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded).and 5 candidates are in advance state of work. He has also guided 28 MTech. Students for dissertation. He has also supervised 2 foreign students for internship. Dr.Bhasker Pant has more than 70 research publication in National and international Journals. He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.